Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers

Year: 2023 | Citations: 455 | Authors: C. Gao, F. Howard, N. Markov, E. Dyer, S. Ramesh

Abstract

Large language models such as ChatGPT can produce increasingly realistic text, with unknown information on the accuracy and integrity of using these models in scientific writing. We gathered fifth research abstracts from five high-impact factor medical journals and asked ChatGPT to generate research abstracts based on their titles and journals. Most generated abstracts were detected using an AI output detector, ‘GPT-2 Output Detector’, with % ‘fake’ scores (higher meaning more likely to be generated) of median [interquartile range] of 99.98% ‘fake’ [12.73%, 99.98%] compared with median 0.02% [IQR 0.02%, 0.09%] for the original abstracts. The AUROC of the AI output detector was 0.94. Generated abstracts scored lower than original abstracts when run through a plagiarism detector website and iThenticate (higher scores meaning more matching text found). When given a mixture of original and general abstracts, blinded human reviewers correctly identified 68% of generated abstracts as being generated by ChatGPT, but incorrectly identified 14% of original abstracts as being generated. Reviewers indicated that it was surprisingly difficult to differentiate between the two, though abstracts they suspected were generated were vaguer and more formulaic. ChatGPT writes believable scientific abstracts, though with completely generated data. Depending on publisher-specific guidelines, AI output detectors may serve as an editorial tool to help maintain scientific standards. The boundaries of ethical and acceptable use of large language models to help scientific writing are still being discussed, and different journals and conferences are adopting varying policies.